Adding OAI-ORE Support to Repository Platforms

Alexey Maslov, Adam Mikeal, Scott Phillips,
John Leggett, Mark McFarland

Texas Digital Library
OR'09

Overview

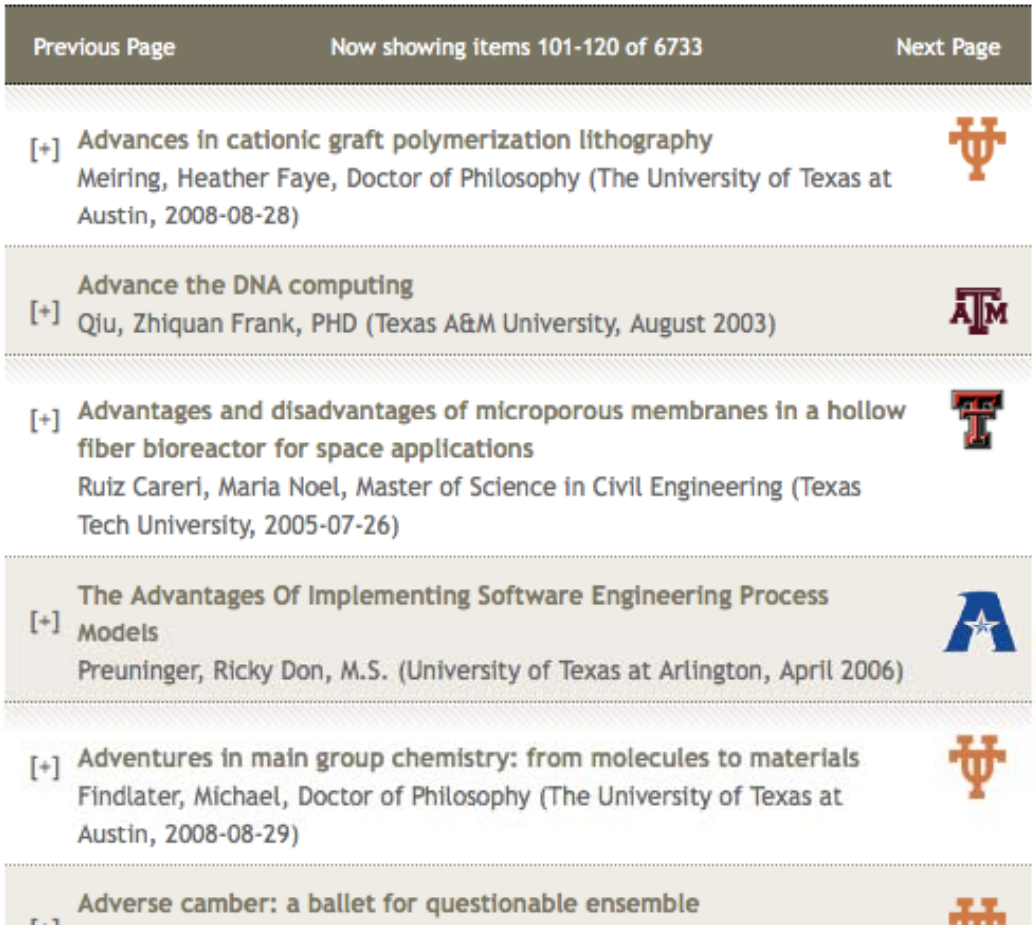
- *Texas Digital Library Use Case for OAI-ORE*
- *Mapping ORE model to DSpace architecture*
- *Implementation*
- *Results and Implications*

Texas Digital Library

- State-wide initiative
- Eighteen members
  - Public/Private
  - Small/Medium/Large

Electronic Theses and Dissertations

- Federated Collection
- Built on top of DSpace/Manakin



Current Federation Method

- Performed via scripted ingest process
- New batch every semester
- Manual corrections to existing content

Replacement Requirements

- Perform maintenance automatically
- Detect changes in existing content
- Support interchange of metadata and content

Harvesting Solution

- Use the Open Archives Initiative Protocol for Metadata Harvesting
- Member institutions as data providers
- TDL Federated Repository as a service provider

* Open Archives Initiative Protocol for Metadata Harvesting
http://www.openarchives.org/pmh/

OAI-PMH, advantages

- Ubiquitous
- Supports selective harvesting
- Tracks changes
- Can be automated

OAI-PMH, obstacles

- No existing harvesting solution for DSpace
- Supports harvesting of metadata specifically

Disseminating content

- How do you disseminate content through a metadata harvesting protocol?
  - Wrap it in a packaging format
  - Include the metadata
  - Encode the references to the files
  - Harvest the package

METS, advantages

- Metadata Encoding and Transmission Standard
- Maintained by the Library of Congress
- Mature standard
- Widely adopted

* Metadata Encoding and Transmission Standard, Library of Congress
http://www.loc.gov/standards/mets/

Packaging, disadvantages

- Complete packaging format
- Open to interpretation
- Ambiguities at the OAI-PMH layer

OAI-ORE

“Open Archives Initiative Object Reuse and Exchange defines standards for the description and exchange of aggregations of Web resources.”

- Specialized
- Simple

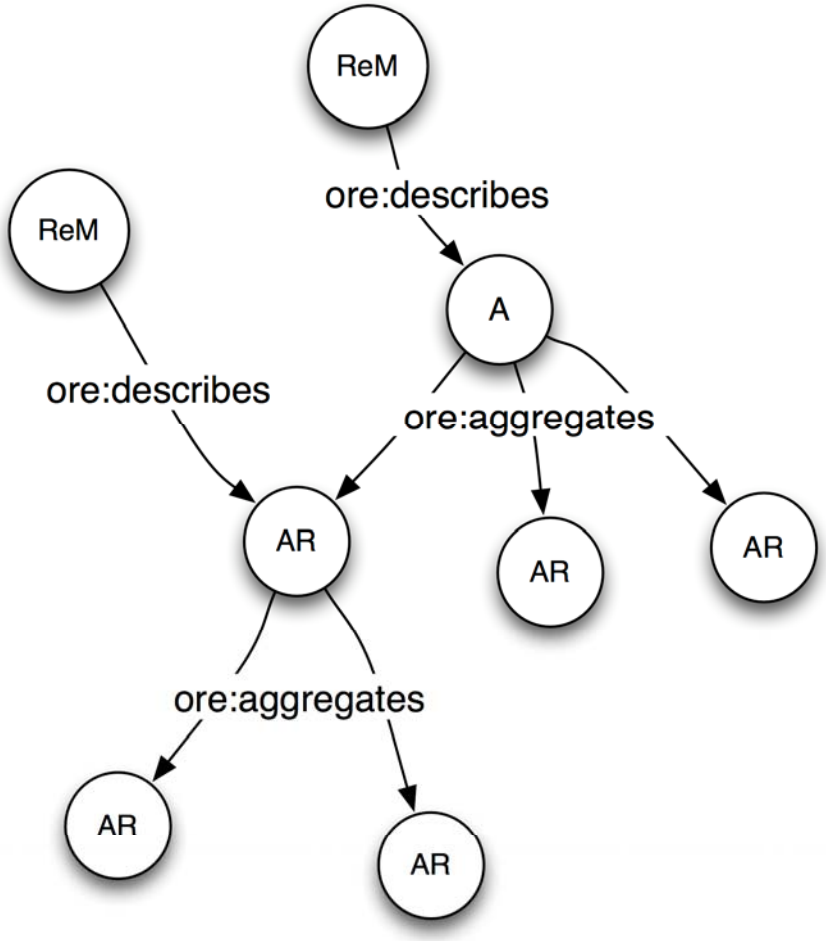
* Open Archives Initiative Object Reuse and Exchange
http://www.openarchives.org/ore/

Mapping DSpace to OAI-ORE

- *ORE Abstract Data Model*
- *DSpace architecture*
- *The Mapping*

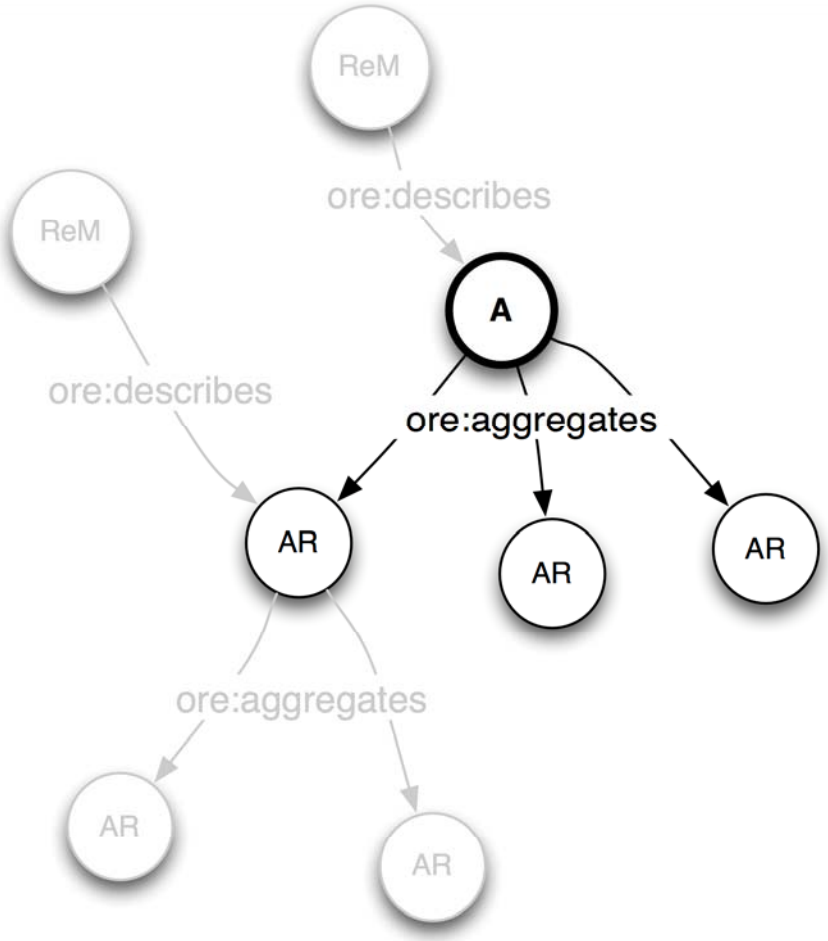
ORE Data Model

- Aggregations
- Aggregated Resources
- Resource Maps



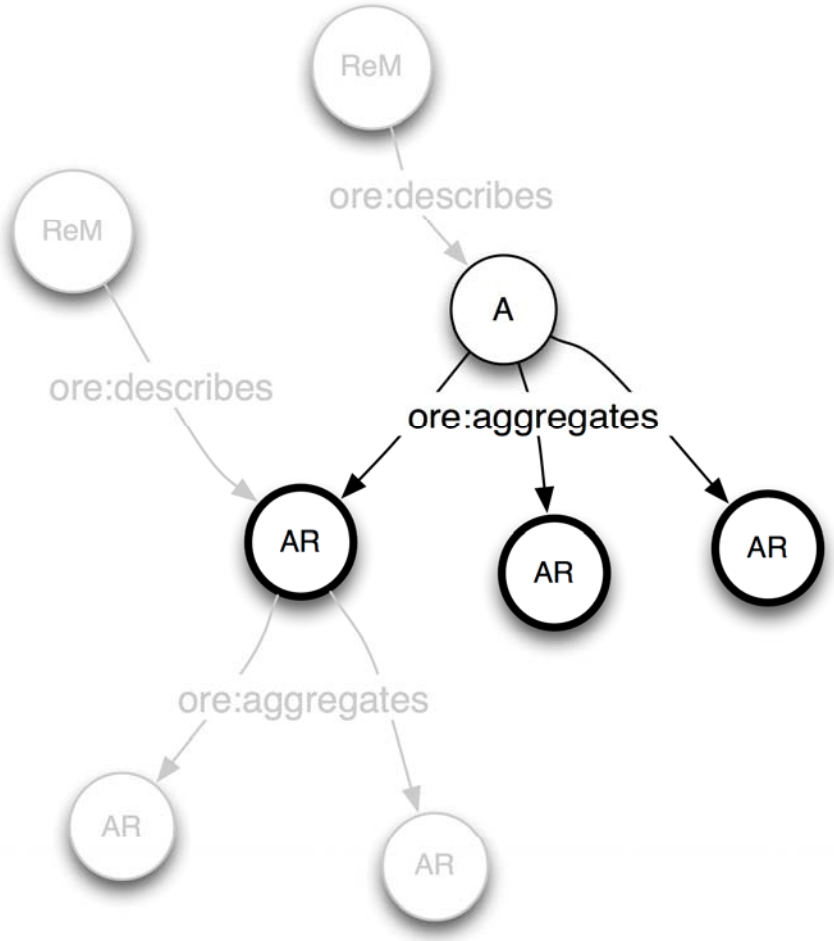
Aggregation (A)

- Describes a set of resources
- Conceptual construct



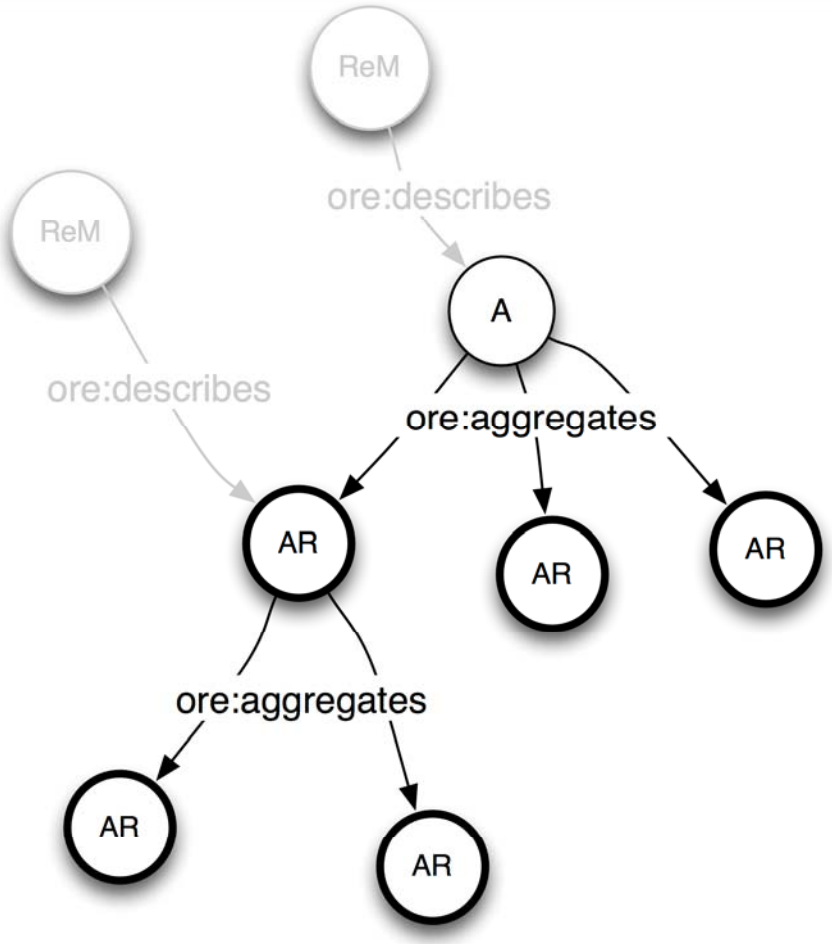
Aggregated Resource (AR)

- Object of interest
- Part of an aggregation
- Can itself be an aggregation



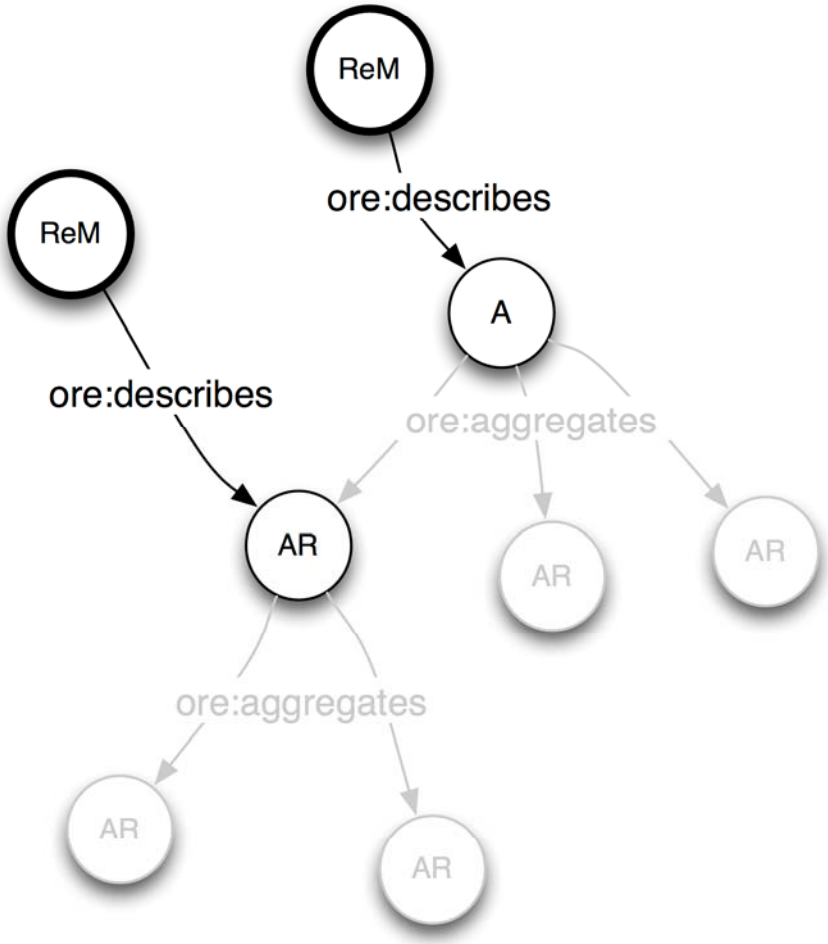
Aggregated Resource (AR)

- Object of interest
- Part of an aggregation
- Can itself be an aggregation



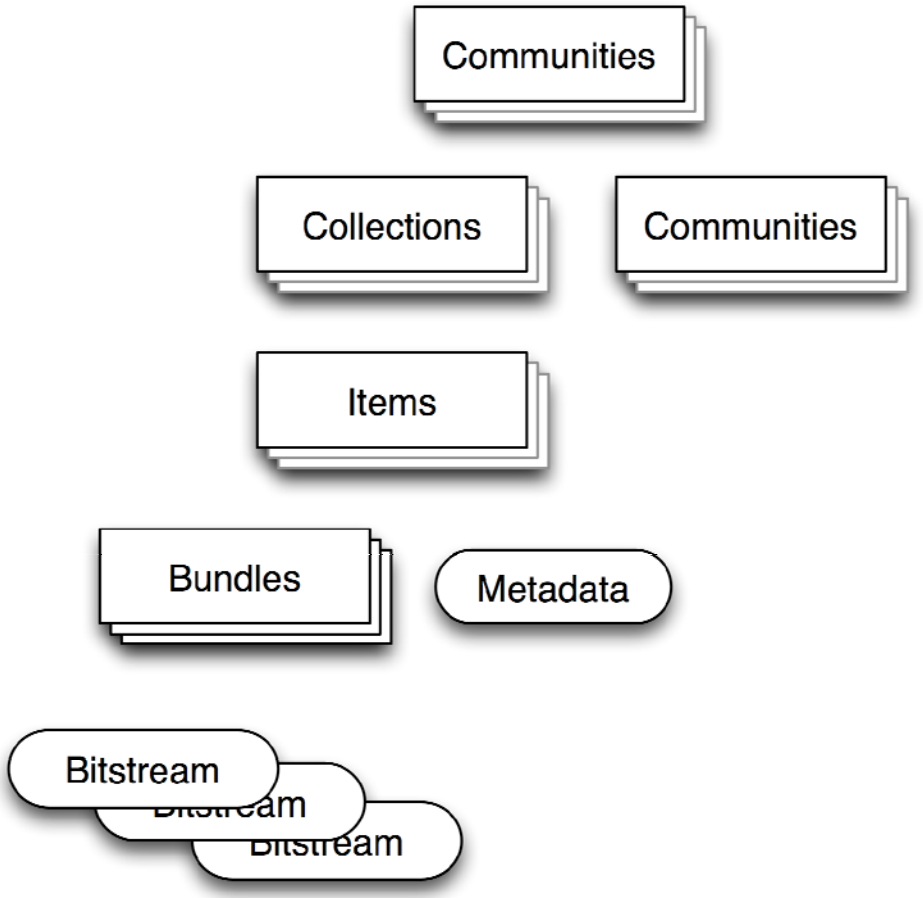
Resource Map (ReM)

- Describes an aggregation
- Enumerates its aggregated resources
- Can be serialized in RDF or Atom XML



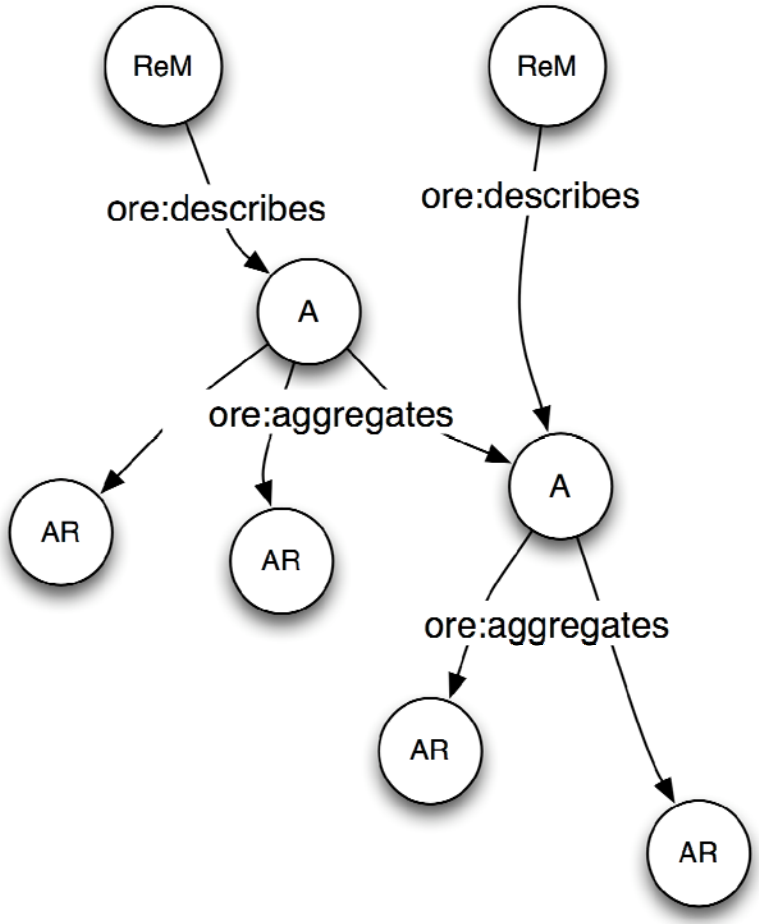
DSpace Model v1.x

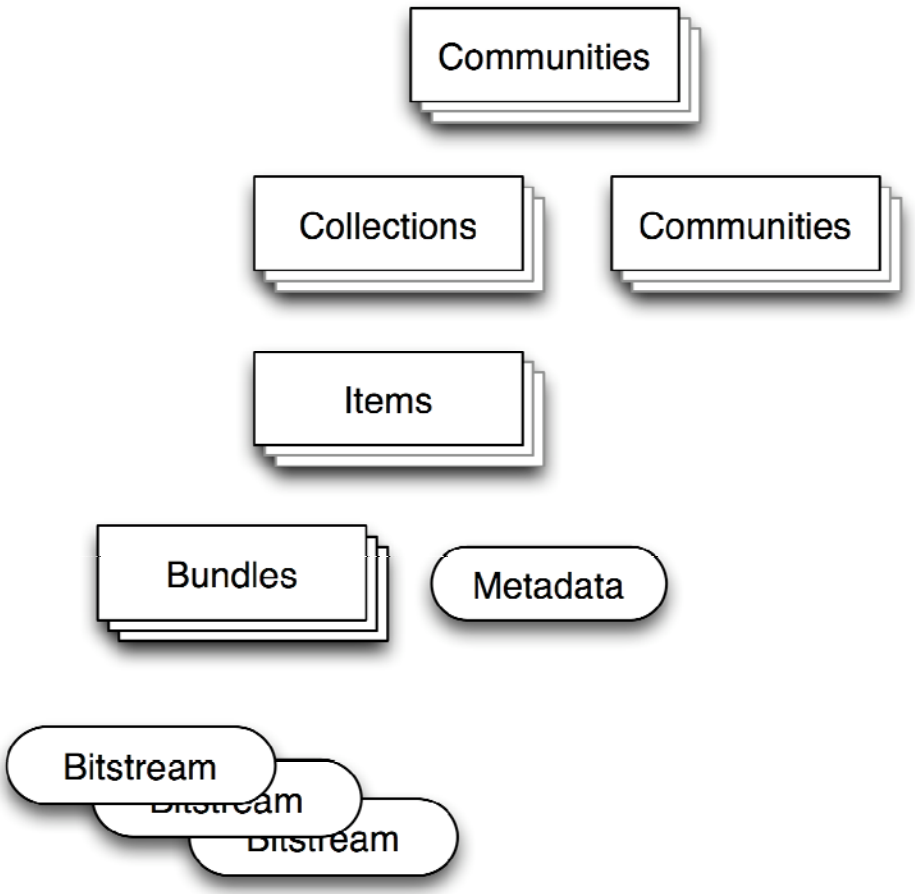
- Communities
- Collections
- Items
- Bundles
- Bitstreams



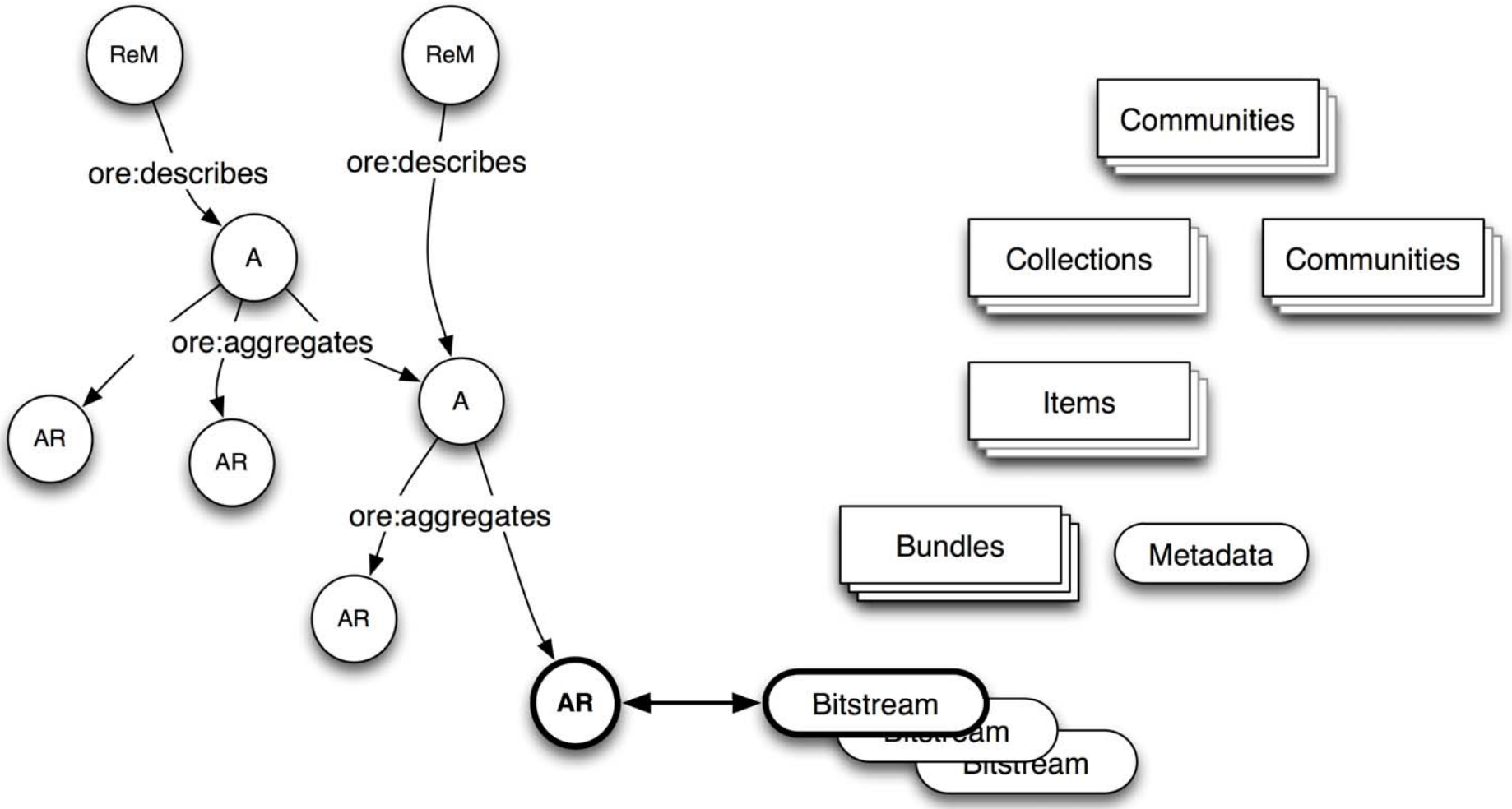
ORE

DSpace

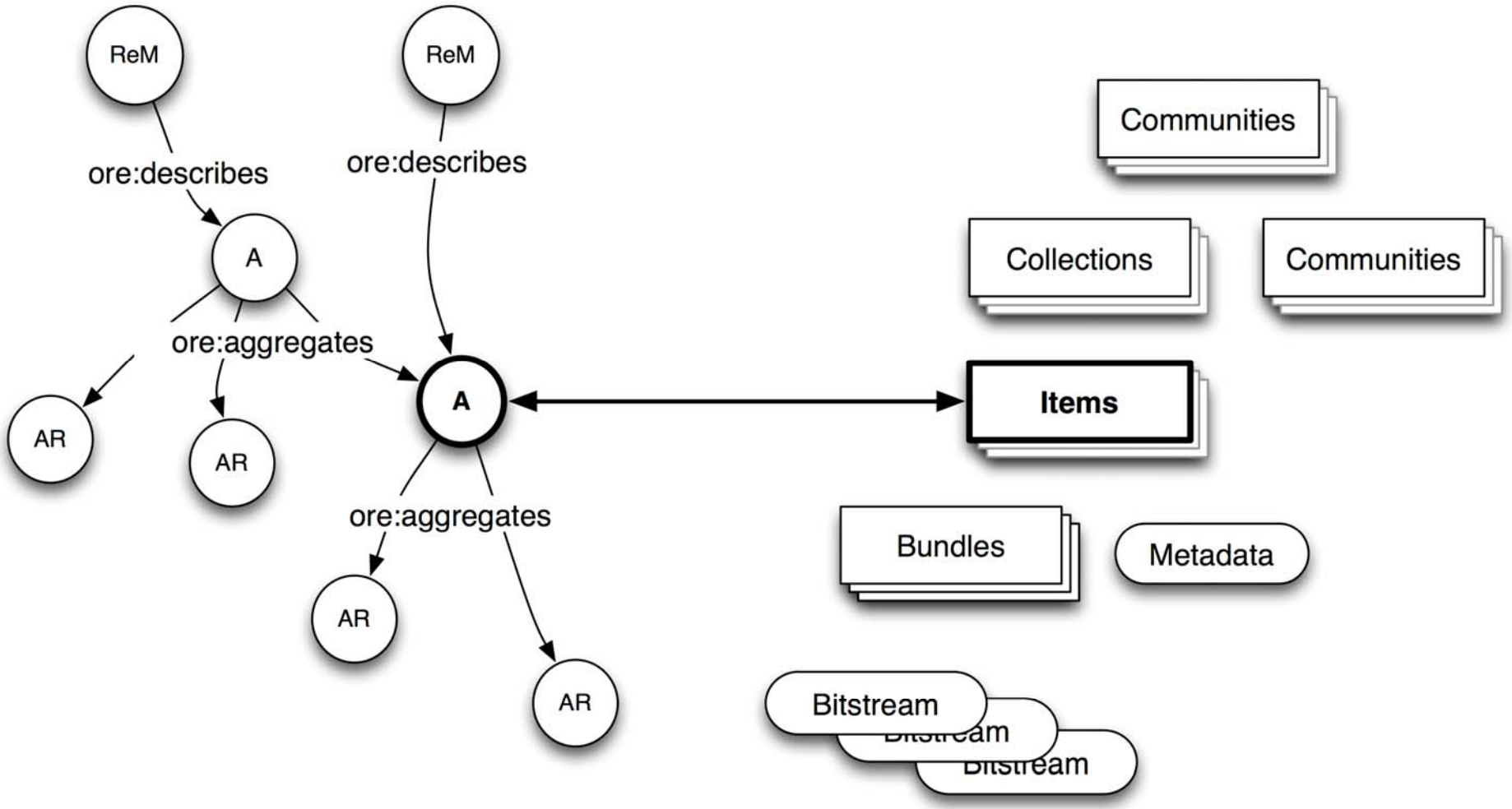

ReM
ReM
ore:describes
ore:describes
A
ore:aggregates
A
AR
AR
ore:aggregates
AR
AR


Communities
Collections
Communities
Items
Bundles
Metadata
Bitstream

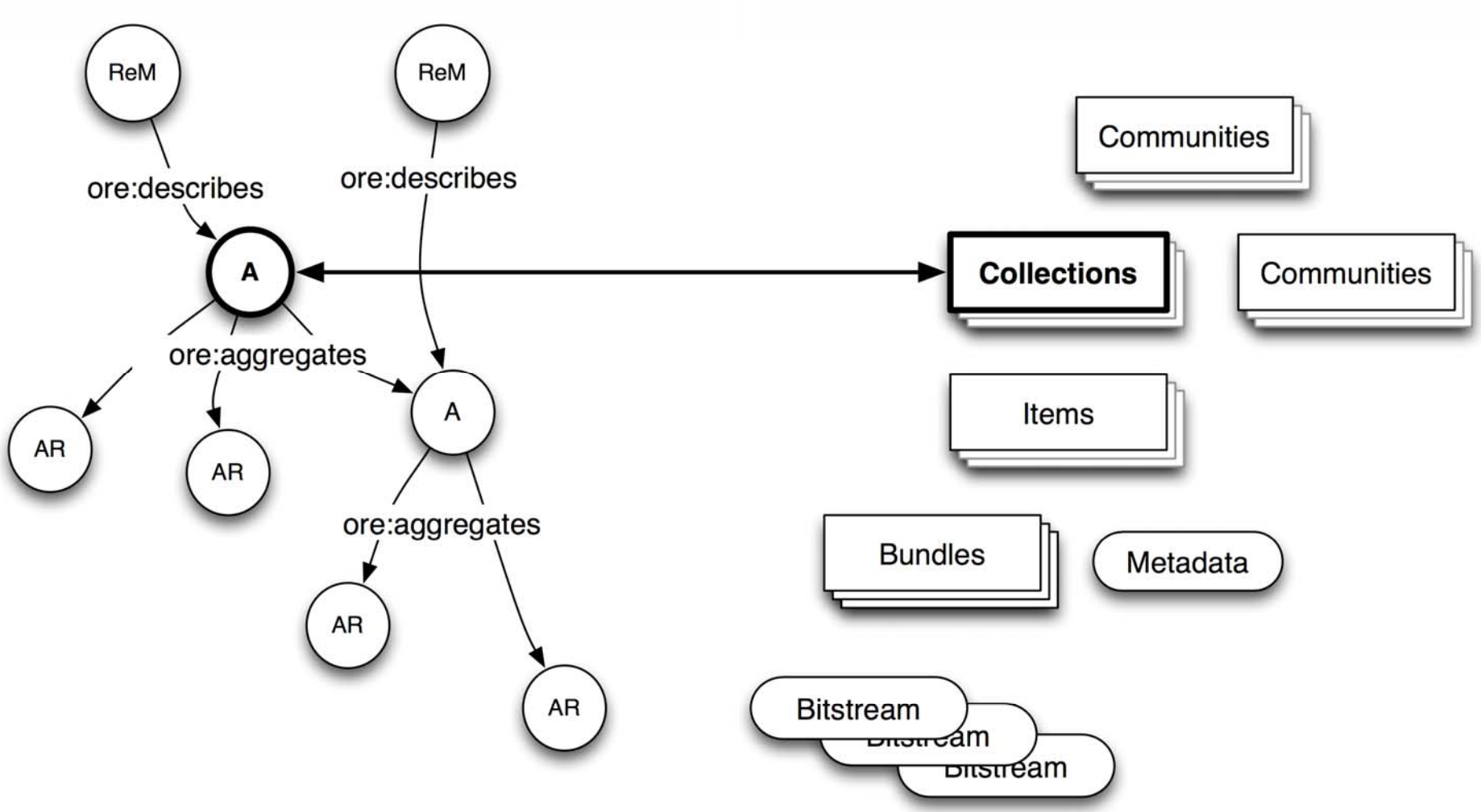
Mapping


ReM
ReM
ore:describes
ore:describes
A
ore:aggregates
AR
AR
A
ore:aggregates
AR
AR
Communities
Collections
Communities
Items
Bundles
Metadata
Bitstream

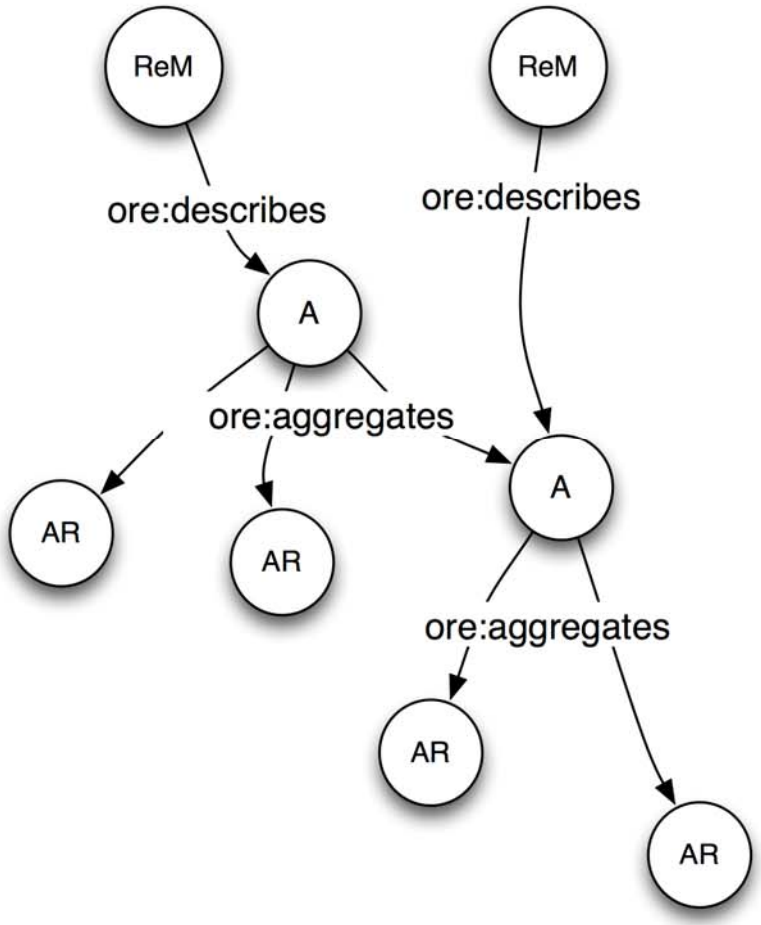
Mapping

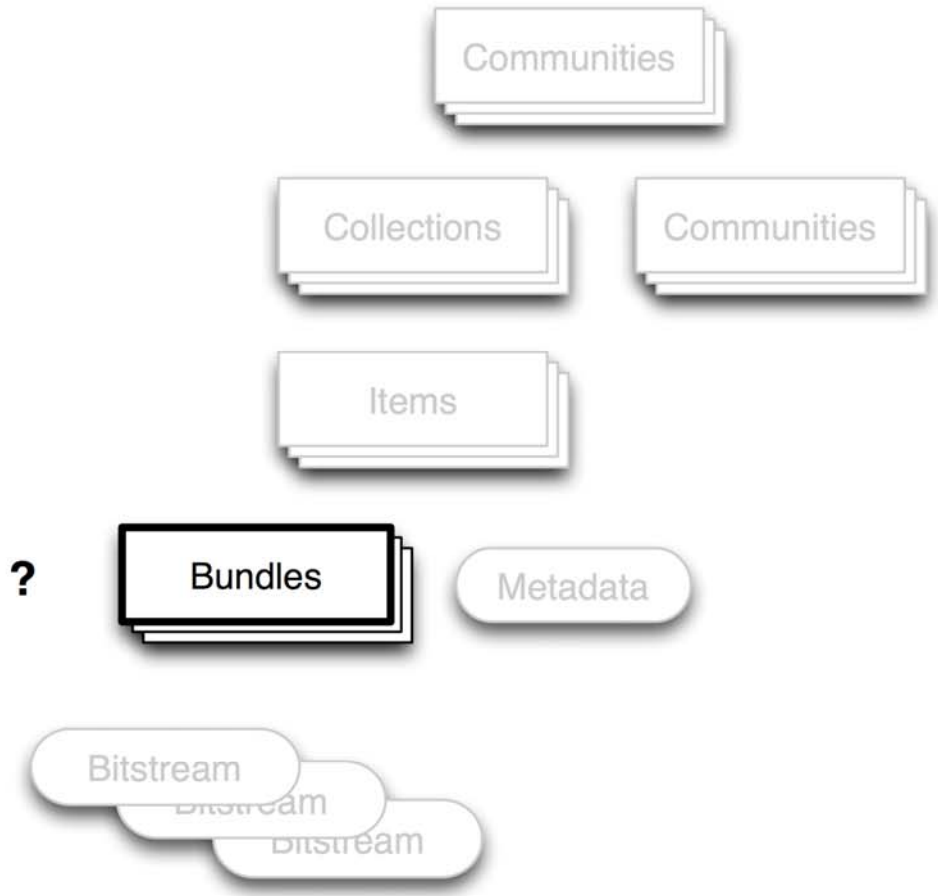

ReM
ReM
ore:describes
ore:describes
A
ore:aggregates
AR
AR
A
ore:aggregates
AR
AR
Communities
Collections
Communities
Items
Bundles
Metadata
Bitstream

Mapping


ReM
ReM
ore:describes
ore:describes
A
ore:aggregates
AR
AR
A
ore:aggregates
AR
AR
Communities
Collections
Communities
Items
Bundles
Metadata
Bitstream

Bundles?


ReM
ReM
ore:describes
ore:describes
A
ore:aggregates
AR
AR
A
ore:aggregates
AR
AR


Communities
Collections
Communities
Items
?
Bundles
Metadata
Bitstream

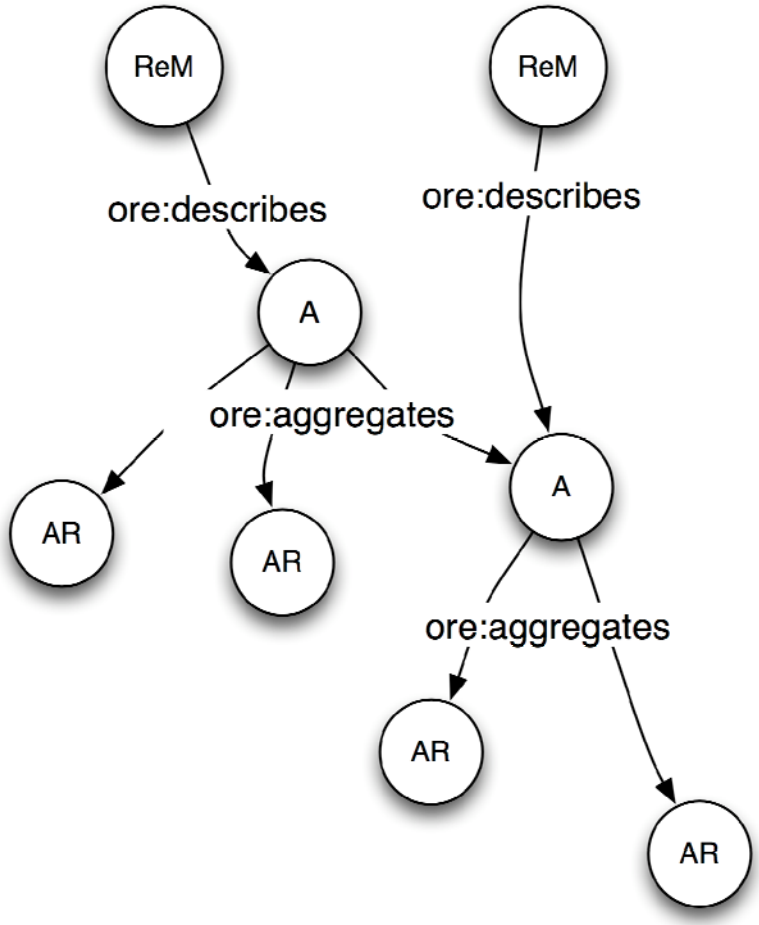
Bundles, Potential Options

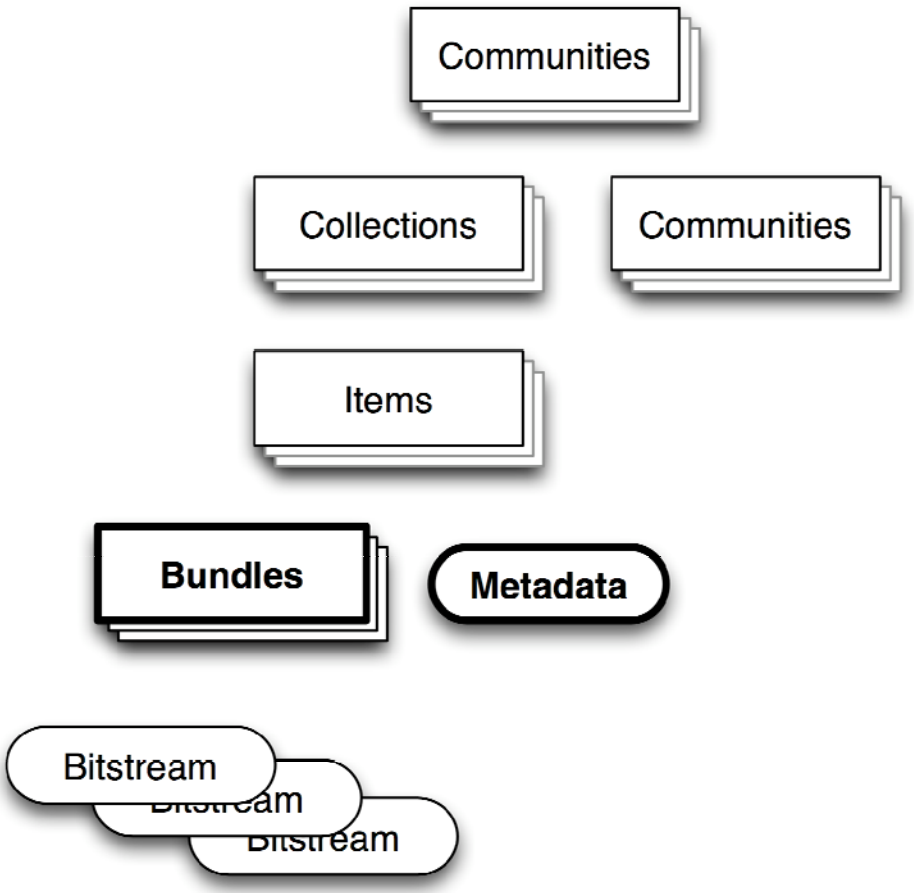
- Bundles as Aggregations of Bitstreams
- Bundles as filters for Aggregated Resources
- Bundles as DSpace-specific metadata

Bundles, Observations

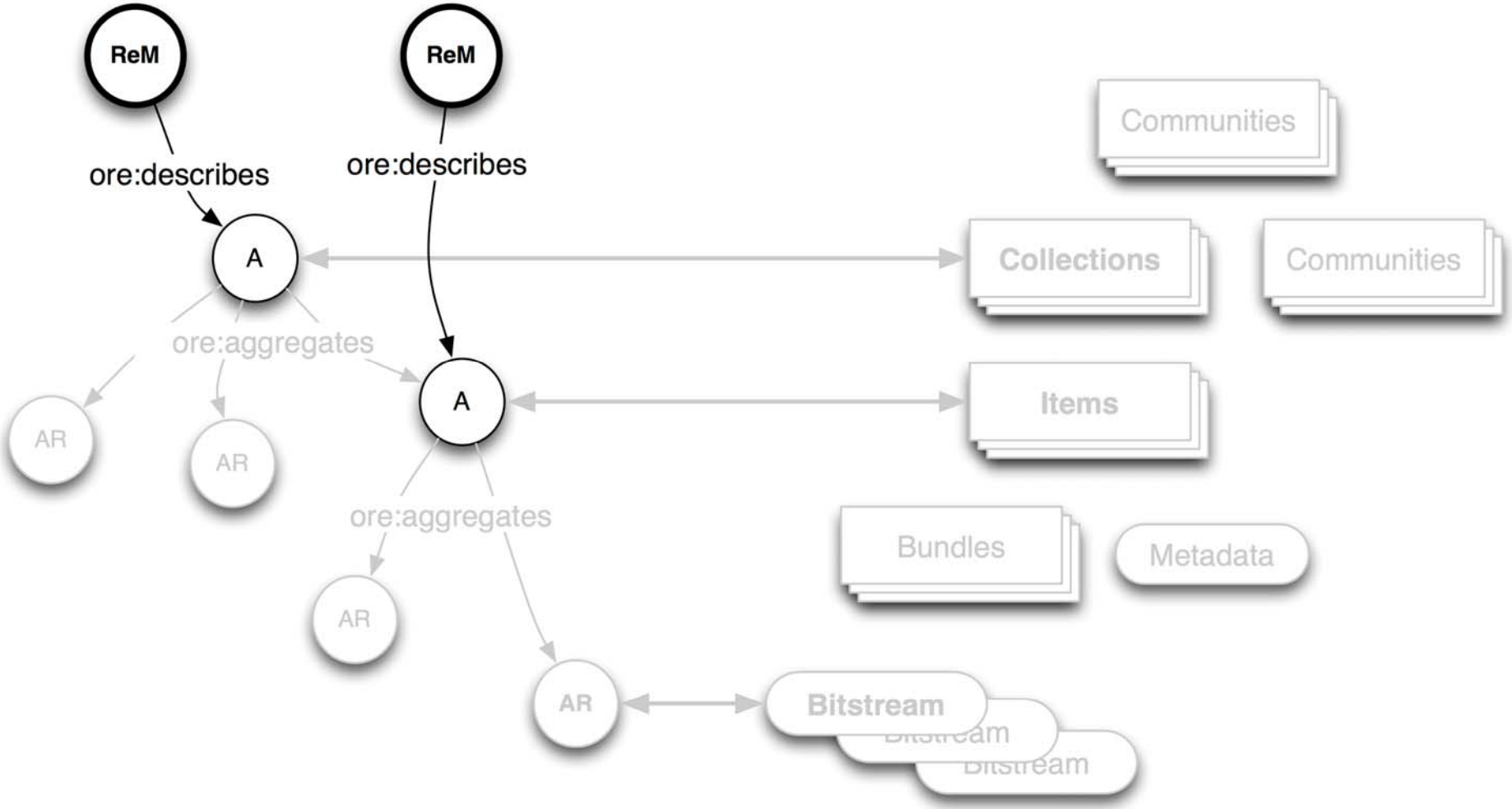
- By default, specialized for internal tasks
- Extendible for any use
- Obscured from the end user

DSpace Bundles





Serialization in Atom



Implementation

- *ORE Dissemination*
- *ORE Harvesting*
- *Automation*

Interfacing with DSpace

- Web UI
- LNI and SWORD
- Ingest and export scripts
- **Crosswalks**
  - **Ingestion**
  - **Dissemination**

ORE Dissemination Crosswalk

- Requires:
  - A DSpace Item
- Produces:
  - Atom-serialized ORE ReM

ORE Dissemination via OAI-PMH

- Dissemination crosswalk produces ORE ReMs from DSpace Items
- OAI-PMH provider disseminates them

ORE Harvesting

- Item-level ORE ReM interpreter
- Collection-level OAI-PMH harvester
- Repository level harvest scheduler

ORE Ingestion Crosswalk

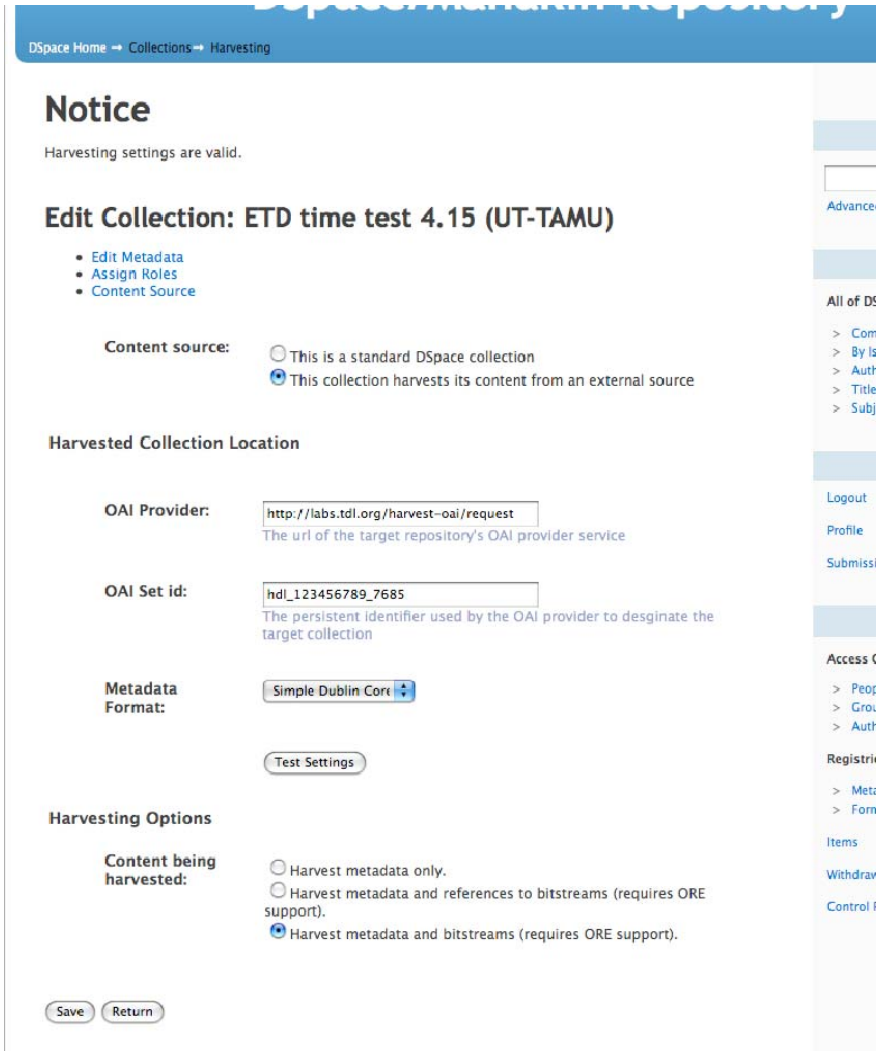
- Requires:
  - A DSpace Item
  - Atom-serialized ORE ReM
- Produces:
  - A DSpace Item with Bitstreams created from AR’s

OAI-PMH Harvester

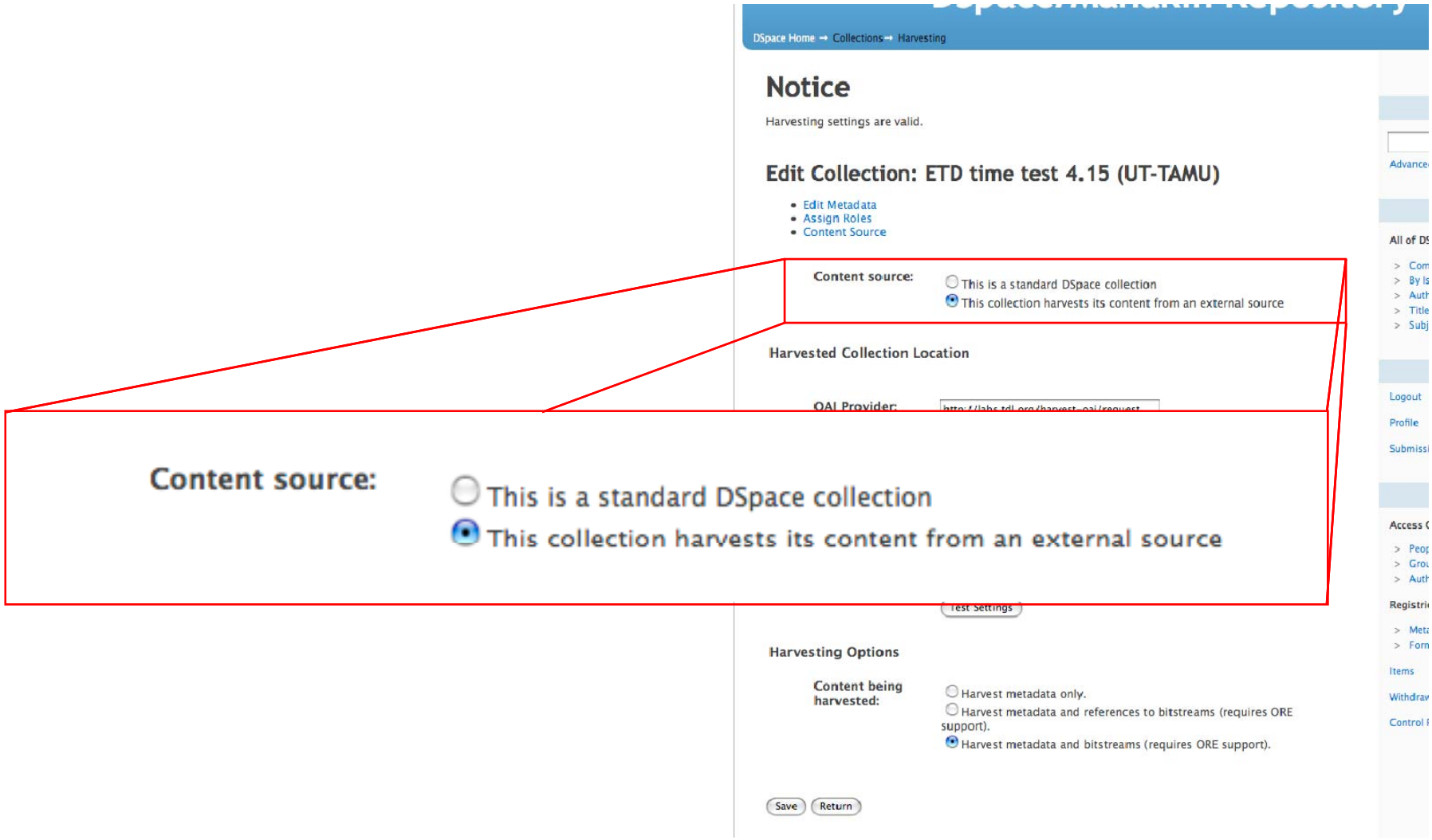
- Queries remote OAI-PMH providers
- Processes responses as individual records
- Implemented at Collection level

Collection Settings

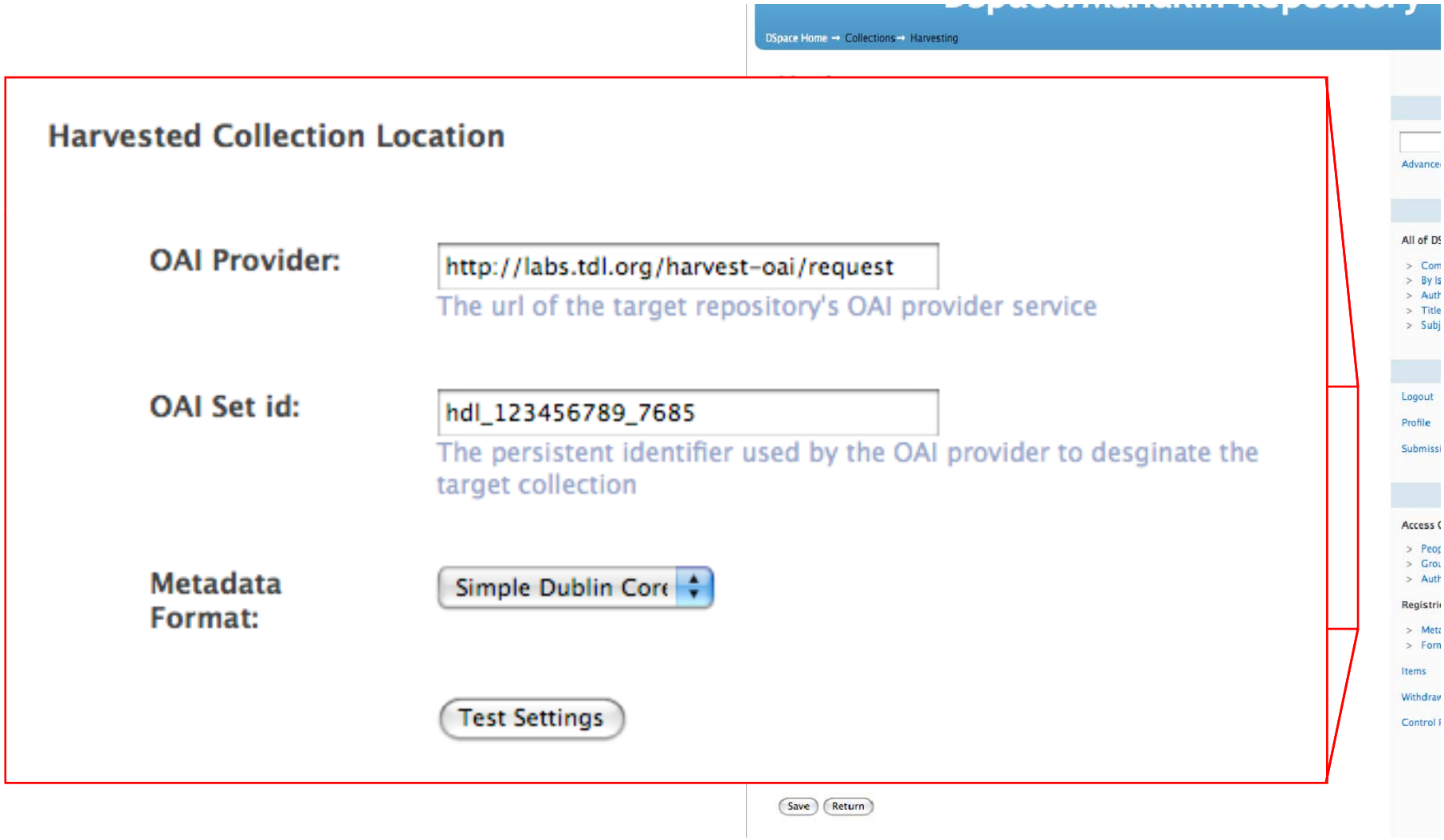
- Source of collection's content
- OAI-PMH provider information
- Harvesting Level



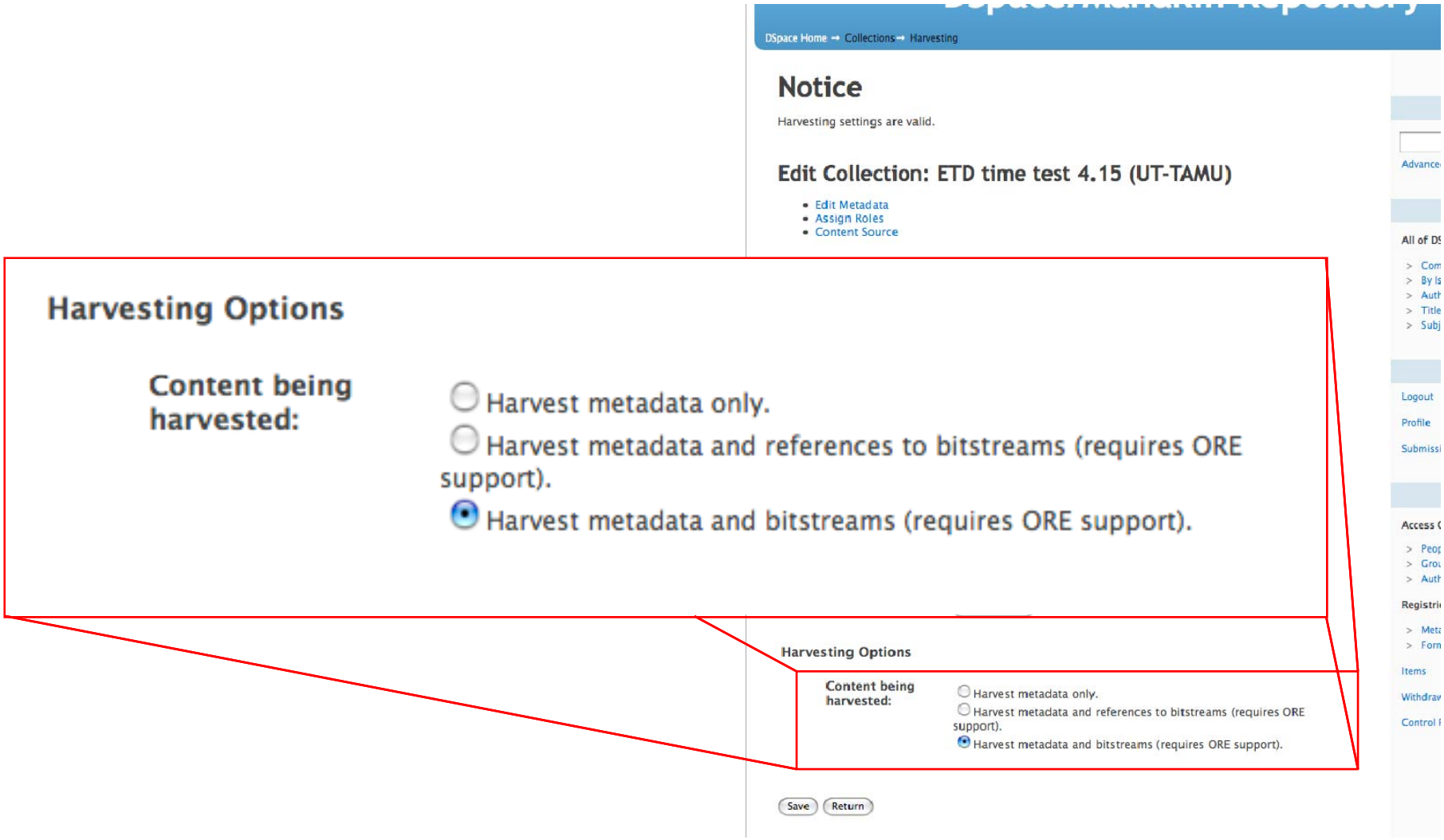
Collection Source



OAI-PMH Settings



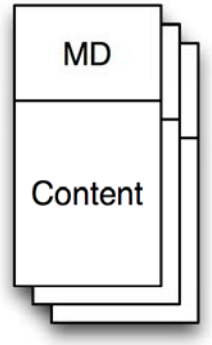
Harvest Level



Harvesting a Collection

Local collection
(OAI-PMH harvester)

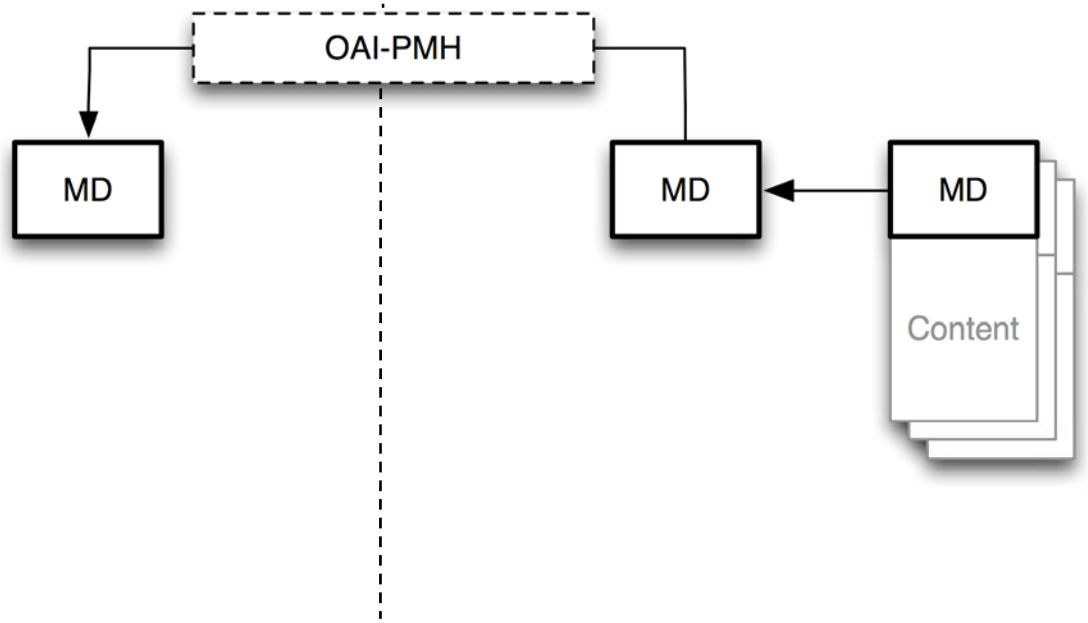
Remote collection
(OAI-PMH provider)



Harvest Metadata

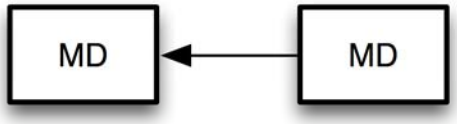
Local collection
(OAI-PMH harvester)

Remote collection
(OAI-PMH provider)

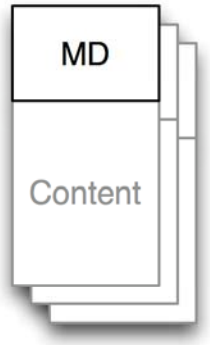


Metadata Replicated

Local collection
(OAI-PMH harvester)

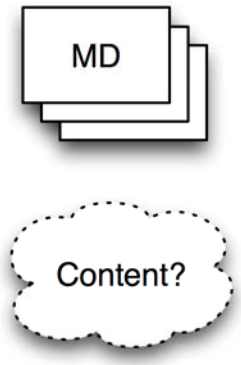


Remote collection
(OAI-PMH provider)

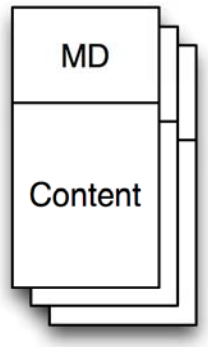


Case 1: Metadata Only

Local collection
(OAI-PMH harvester)



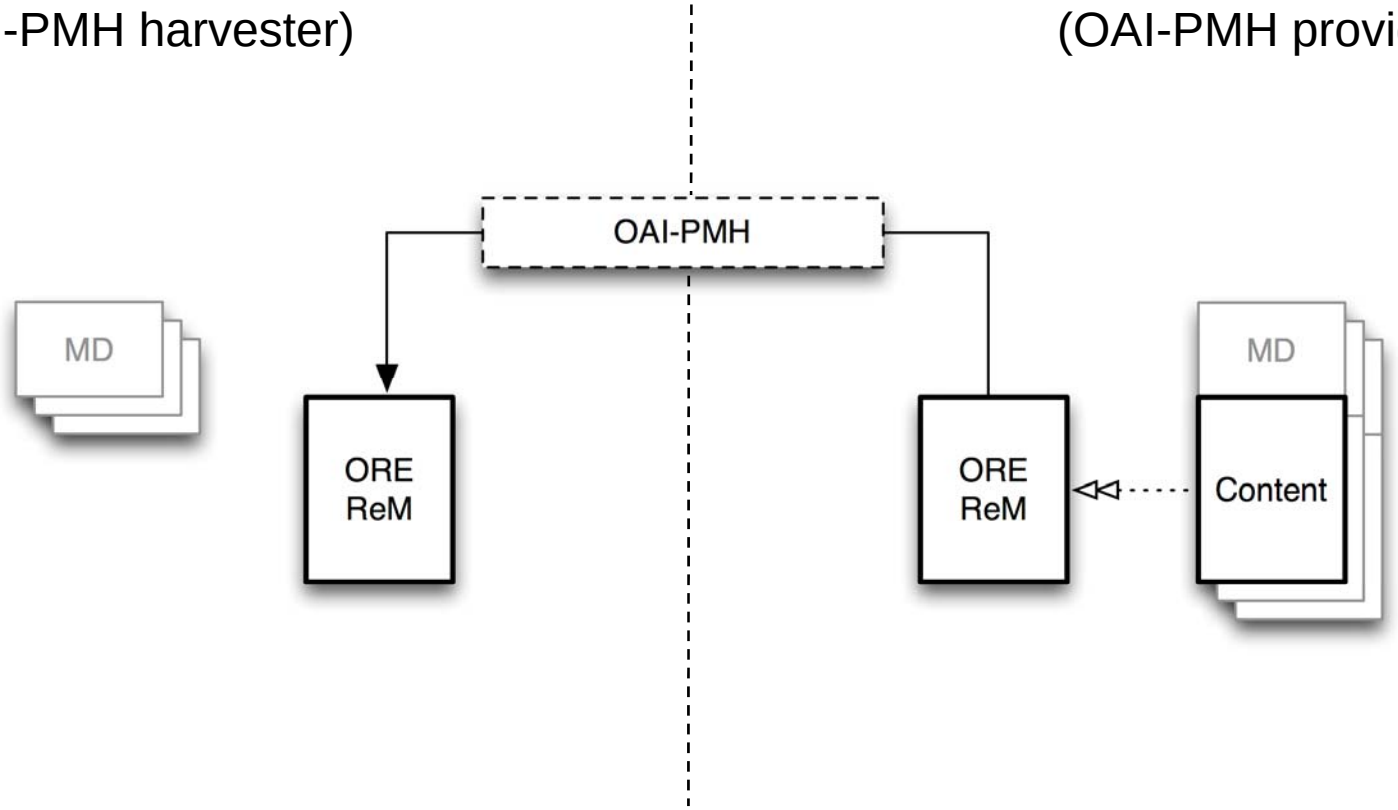
Remote collection
(OAI-PMH provider)



Harvest ORE ReMs

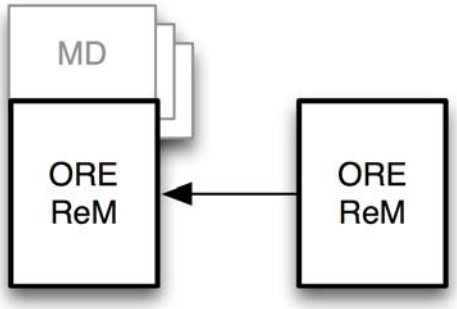
Local collection
(OAI-PMH harvester)

Remote collection
(OAI-PMH provider)

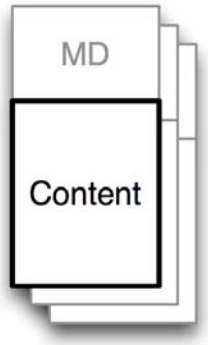


Case 2: Metadata + Content Ref's

Local collection
(OAI-PMH harvester)



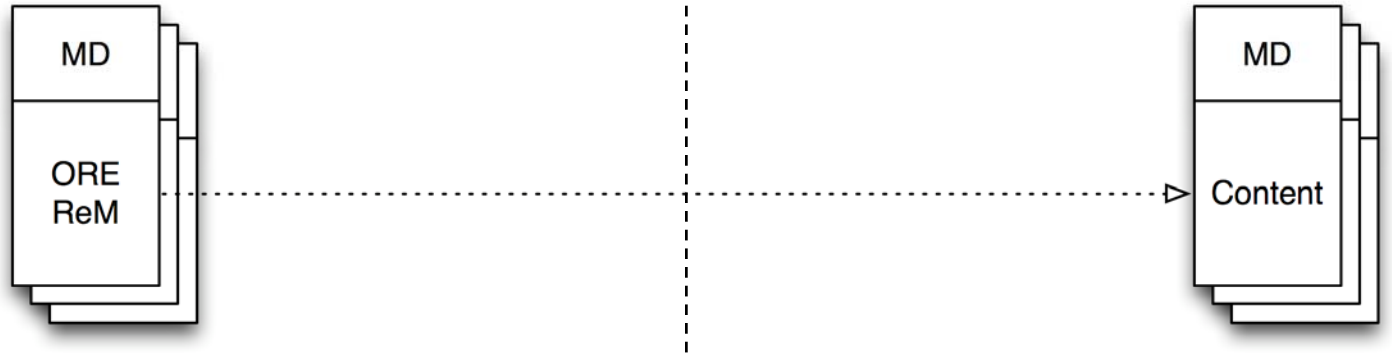
Remote collection
(OAI-PMH provider)



Case 2: Metadata + Content Ref's

Local collection
(OAI-PMH harvester)

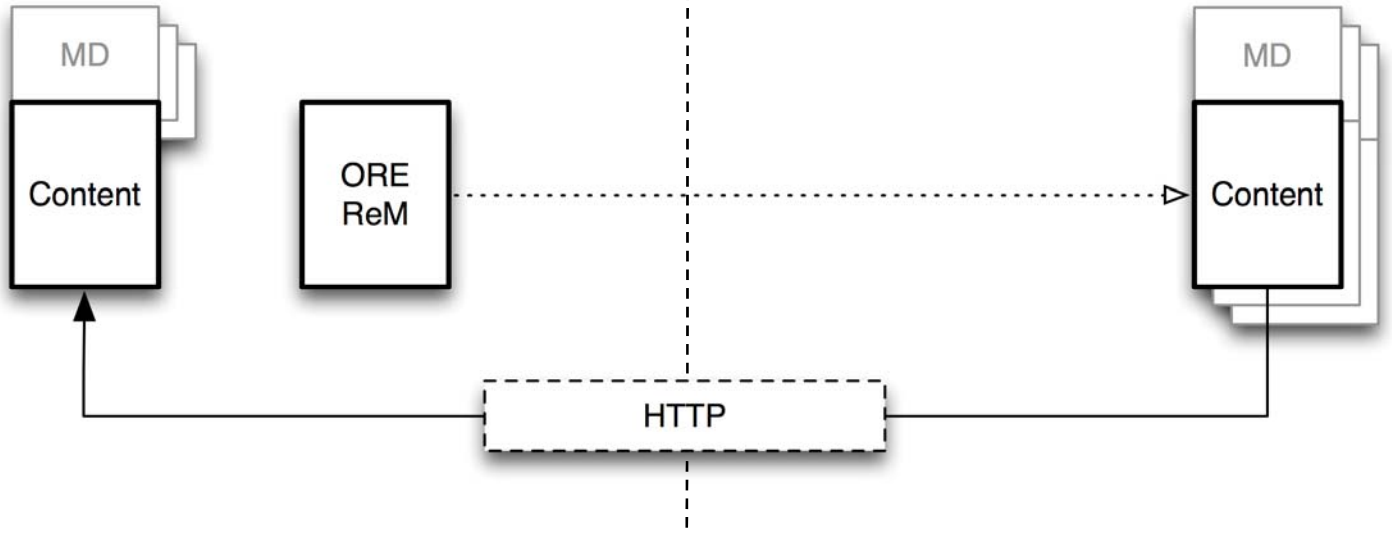
Remote collection
(OAI-PMH provider)



Case 3: Metadata + Content

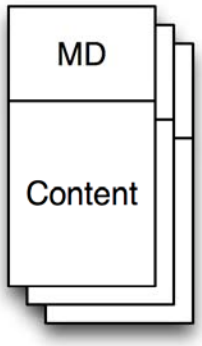
Local collection
(OAI-PMH harvester)

Remote collection
(OAI-PMH provider)

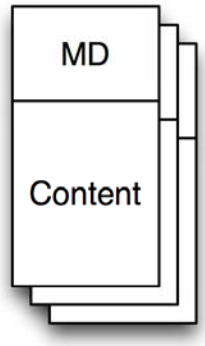


Case 3: Metadata + Content

Local collection
(OAI-PMH harvester)



Remote collection
(OAI-PMH provider)



Harvest Scheduling System

- Monitors harvested collections
- Starts harvests at regular intervals
- Alerts administrators of errors

Results

- *The Primary Use Case*
- *TDL in General*
- *The Greater Web Community*

Harvesting using PMH+ORE

- Federated ETD collection currently in pre-production at TDL
- Addresses primary requirements
  - Performs maintenance automatically
  - Detects changes in existing content
  - Supports interchange of metadata and content

Other Possibilities

- Specialized DSpace instances
- Flexible repository architecture
- Interoperability with other repository systems

"Large-scale ETD repositories: A case study of a digital library application",
Adam Mikeal, James Creel, Alexey Maslov, Scott Phillips, John Leggett, Mark McFarland.
JCDL 2009

Current Priorities

- Live deployment at TDL
- Release to the open source community
- Integration into DSpace 1.6

INSTITUTE *of*
Museum and Library
SERVICES

National Leadership Grant #LG-05-07-0095-07

Questions?